Санкт-Петербургский государственный университет

Факультет прикладной математики процессов управления
Кафедра технологии программирования

Ахремчик Ян Валерьевич

# Распознавание и извлечение 3D-моделей по двумерным изображениям

Выпускная квалификационная работа

Направление 01.03.02

Прикладная математика, фундаментальная информатика и программирование

Научный руководитель:
ст. преподаватель Стученков А. С.

Санкт-Петербург
2019

# Оглавление

# Введение

Множество объектов окружают человека в реальном мире. У них разнятся форма, структура, цвет, размер. И, хотя, человек умеет взаимодействовать с всеми различными типами объектов, современные роботизированные системы весьма ограничены в этом плане. У роботизированных систем существует четкий набор инструкций при работе с предметами той или иной формы. Этим набором инструкций и ограничивается область применения конкретной системы. Умение же анализировать объект позволило бы расширить область применения той или иной роботизированной системы. Более того, так как взаимодействие с объектом происходит в трёхмерном пространстве, то и анализировать форму тоже необходимо в трёх измерениях.

Направление в компьютерном зрении, которое связано с этой задачей называется "Трёхмерная реконструкция". Вообще говоря, область применения решений в данном направлении гораздо шире, нежели взаимодействие роботизированных систем с реальным миром. Как пример, можно рассмотреть задачу взаимодействия с предметами в дополненной реальности, трёхмерную реконструкцию человеческого тела, детальную оценку дорожной ситуации.

Есть множество способов реконструкции трёхмерных моделей и они в корне различаются. Выделяют два основных набора методов при реконструкции: активные и пассивные.

Активные методы подразумевают исследование объектов с помощью использования определённого излучения, направленного на объект, а затем считывания данных, отразившихся от объекта. Например: структурное освещение, лазерные дальномеры, лидары, радиоизлучения, ультразвуковые волны, микроволновые излучения и так далее. Однако для этого необходимо специфическое оборудование, что, несомненно, является минусом данного подхода.

Пассивные же методы не производят никакого воздействия на объект, они лишь используют набор датчиков для измерения естественного излучения, отражаемого объектом. Типичным примером являются

матрицы камер. При использовании камер выделяют бинокулярные и монокулярные схемы.

В первом случае используется стереопара из двух камер. Используя два ракурса, строится карта глубины снятой сцены [9]. Имея карту глубины, получается 2.5 мерное пространство. Используя 2.5 мерное представление объекта, с помощью методов глубокого обучения получается полноценная трёхмерная модель [6][12].

Во втором случае, при использовании одной камеры, принцип реконструкции состоит в том, чтобы отснять набор кадров объекта с разных ракурсов, либо заснять интересующий объект на видео, чтобы в дальнейшем реконструировать объект из набора изображений[20]. Относительно новым подходом является реконструкция объекта по единственному монокулярному изображению.

Существует множество решений, позволяющих реконструировать трёхмерную модель по одному изображению, однако ни одно из этих решений не реализует принцип, позволяющий извлекать несколько трёхмерных моделей из одного изображения реального мира.

# Постановка задачи

## Цель

Целью данной выпускной квалификационной работы является разработка решения, позволяющего осуществлять процесс извлечения набора трёхмерных изображений из фотографий реального мира.

## Задачи

- Провести теоретический анализ предметной области.
- Провести анализ и сравнение существующих решений.
- Разработать решение.
- Произвести тестирование разработанного решения.

# Обзор литературы

1. Hartley R. I., Zisserman A. Multiple View Geometry in ComputerVision

В этой книге рассматриваются проблемы компьютерного зрения по интерпретации изображения объекта, сделанного с разных точек обзора. И, хотя по современным меркам она довольно старая(2003г), основополагающие понятия из неё будут в ходу ещё очень долго.

2. Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, Abhinav Gupta Learning a Predictable and Generative Vector Representation for Objects

Очень хорошая статья, дающая представление о интерпретации трёхмерного объекта в виде вектора. Так же описывает общую структуру подходов по реконструкции трёхмерных изображений. Более того, в ней описываются процессы генерации, а так же вычитания трехмерных изображений одного из другого.

3. Daeyun Shin, Charless C. Fowlkes, Derek Hoiem Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction

Статья использовалась для понимания различий между представлениями трёхмерных моделей.

# 1. Глава 1. Анализ предметной области

Изучая предметную область, следует отталкиваться от типа решаемой задачи. В более общем виде проблема звучит как "Single view reconstruction" (далее SVR) или "реконструкция, с использованием одного вида", (как вид следует понимать изображение).

Анализируя поведение человека, он (человек), как движущийся организм, имеет возможность изучать объекты с множества ракурсов. Эти множества ракурсов дают человеку представление о геометрии изучаемого объекта. В дополнении к тому, что человек может воспринимать трёхмерные объекты, при наблюдении с некоторых точек зрения, он так же способен воспринимать трёхмерную модель объекта, бросив всего лишь один взгляд на него. Уже однажды изученный объект возможно восстановить с одного взгляда, благодаря общим знаниям о его структуре. Для алгоритма восстановить трёхмерную модель объекта по одному изображению - достаточно сложная проблема, ввиду того что количество информации, которое несёт в себе одно изображение, недостаточно для трёхмерной реконструкции. В связи с этим родилось несколько подходов, позволяющих обойти это ограничение.

## 1.1. Форма из текстуры

Существует не так много методов восстановления модели поверхности, с помощью проекции текстурного поля, которое лежит на этой поверхности [23] (рис. 1). Рассматривают два типа методов, глобальные и локальные.

**Глобальные** методы направлены на восстановление целостной модели поверхности, используя предположение о распределении текстурных элементов. Эти предположения называются изотропией или гомогенностью. Методы, базируемые на гомогенности предполагают что текстурные элементы являются результатом "пуассоновского процесса", который утверждает о том, что если набор случайных точек в некотором пространстве образует пуассоновский процесс, то число точек в области конечного размера является случайной величиной с распре-

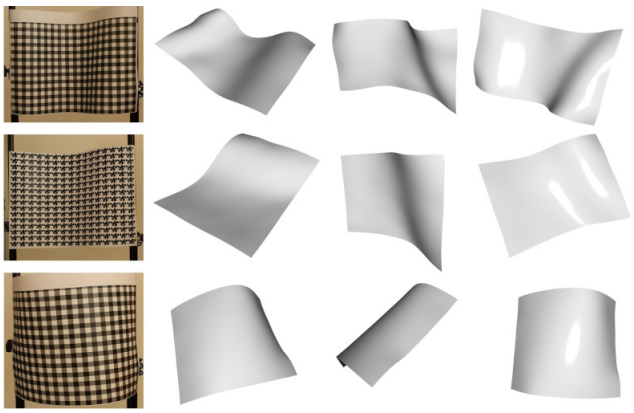

Рис. 1: Форма из текстуры

делением Пуассона. Однако не учитывается факт деформаций отдельных текстурных элементов. Что является главным недостатком данного подхода, так как крайне мало количество гомогенных текстур.

**Локальные** методы восстанавливают дифференциальные геометрические параметры в точке на поверхности (нормаль и кривизна). Однако методы имеют серьёзный недостаток, необходимо знать что координатные рамки текстурного элемента формируют область кадра, которая локально параллельна вокруг рассматриваемой точки, либо знать дифференциальное вращение поля кадра. Ещё одна важная проблема заключается в получении данных. Все эти методы дают локальную оценку нормали и кривизны. Как результат, как одна локальная оценка может быть полезной, нет никаких оснований полагать что семейство таких оценок будет согласующимся.

Методы поверхностной интерполяции, а так же методы реконструкции формы из текстуры, в которых последние играют важную роль, не применяются больше в компьютерном зрении, так как невозможно оценить те участки объекта, в которых данные (текстурные элементы) отсутствуют.

## 1.2. Форма из тени

Подход метода "формы из тени" (Shape From Shading, SFS) заключается в вычислении трёхмерной формы поверхности с использованием яркости одной чёрно-белой фотографии этой поверхности.

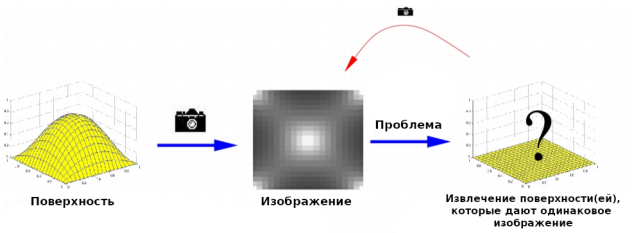


Рис. 2: Форма из тени

На сегодняшний день проблема SFS считается некорректно поставленной. Множество статей показывает, что решение не уникально[5][14]. Очень часто трудности возникают с вогнутыми/выпуклыми поверхностями (рис. 3). На этом изображении продемонстрирована неопределённость, обусловленная оценкой параметров освещения. С одной стороны можно полагать, что изображено два кратера. Однако, если представить источник света находящимся снизу, то изображение кратеров превращается в изображение вулканов(чем оно по сути и является). Данная неопределённость легко обобщается на большое количество примеров.

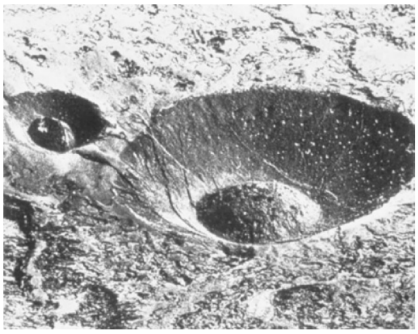

Рис. 3: Неопределённость в виде кратера

В [4] Белхумер с коллегами доказывают, что при условии неопределённости направления освещения и коэффициента отражения Ламберта, одно и то же изображение может быть представлено непрерывным семейством поверхностей. Иными словами, он показывает, что затенение объекта, который рассматривается только с одной точки, не позволяет восстановить его точную трёхмерную модель.

Ввиду вышеизложенных фактов применение SFS для реконструкции трёхмерной модели по единственному изображению не приемлемо.

## 1.3. Глубокое обучение

Ещё одним путём трёхмерной реконструкции является использование знания о представлении объекта и его форме, то есть использование априорной информации о форме объекта. Главным преимуществом использования ”априорных форм” (prior form) является отсутствие необходимости поиска точных соответствий особенностей между изображениями, снятыми с разных ракурсов. В таком случае возможно осуществить трёхмерную реконструкцию по двумерному изображению (рис. 4). В последнее время возросло количество решений, использу-

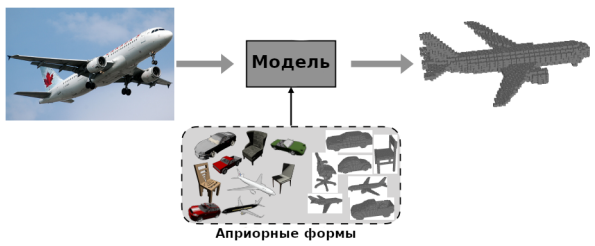


Рис. 4: Подход с использованием априорных форм

ющих данный подход. Причиной этому послужило появление большого количества датасетов с трёхмерными изображениями ShapeNet[18], Pascal3D+[21], ObjectNet3D[13], Pix3d[15], а так же широкое распространение и совершенствование технологий глубинного обучения. Однако этот метод вызывает достаточно большие трудности. Для того, чтобы выучить априорные формы, необходимо большое количество аннотированных трёхмерных объектов. Так как получить качественные аннотированные трёхмерные изображения из реального мира достаточно сложно, большинство подходов используют синтетические данные, полученные процессом рендеринга трёхмерных моделей. Эти подходы используют схожую архитектуру автоэнкодер, так же известную как кодировщик-декодировщик. Кодировщик отображает двумерные изображения в скрытое представление, декодировщик же отображает это представление в трёхмерный объект. (рис. 5) Ввиду обширного количе-

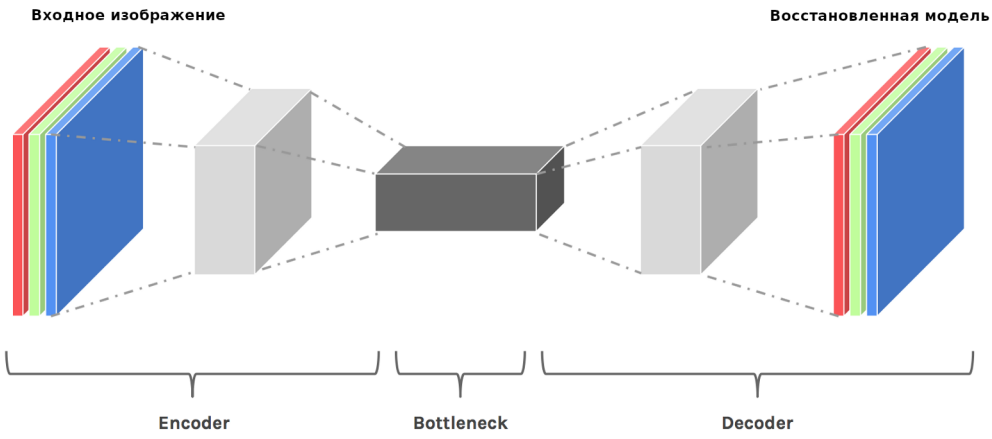


Рис. 5: Пример архитектуры encoder-decoder

ства решений, динамического развития данной области, а так же отсутствия весомых ограничений на условия съёмки было принято решение воспользоваться методами глубокого обучения.

# 2. Глава 2. Анализ и сравнение

## 2.1. Анализ

Решение, выполняющее поставленную задачу должно включать в себя две составляющие:

- Решение, позволяющее производить детектирование и извлечение интересующих объектов из изображений реального мира
- Решение, реконструирующее трёхмерную модель из извлечённого изображения

## 2.2. Детектирование объектов

Для оценки качественной работы алгоритмов будет использоваться метрика средней точности (Average Precision, AP), где точность вычисляется как

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

TP - истинно-положительное решение
FP - ложно-положительное решение

### 2.2.1. Faster R-CNN (Region-based Convolutional Neural Networks)

Faster R-CNN [8] представляет собой усовершенствованную версию архитектуры R-CNN.

**R-CNN** Суть заключается в предсказании регионов, используя процесс, называемый выборочный поиск (Selective search). Данный процесс "смотрит" на изображение через "окна" разных размеров и для каждых размеров пытается сгруппировать пиксели, основываясь на цвете, интенсивности, текстуре для того, чтобы идентифицировать объект. (рис. 6) Как только предполагаемые регионы выделены, R-CNN сжимает регион до размера стандартного квадрата и пропускает через

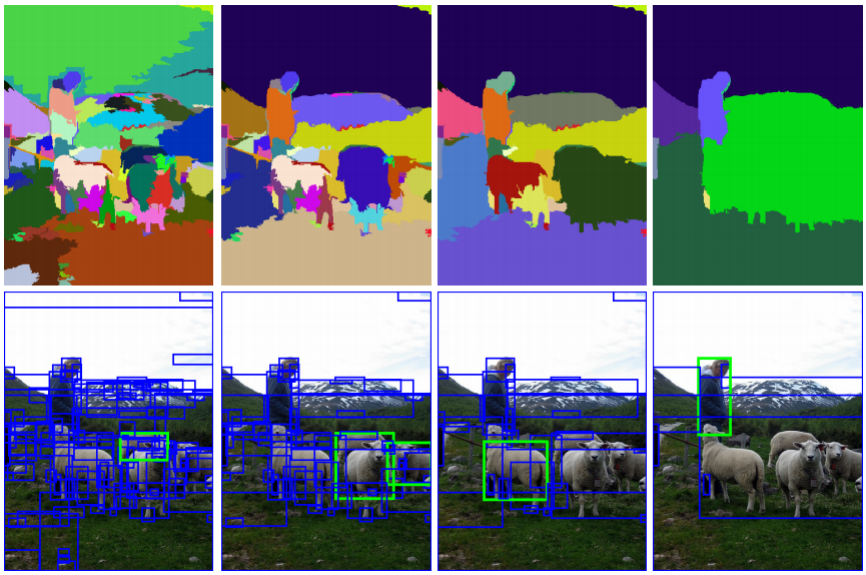

Рис. 6: Принцип работы Selective search

модифицированную версию сети Alexnet. На последнем слое, R-CNN использует метод опорных векторов (SVM), чтобы определить есть ли объект на изображении и классифицировать его. (рис. 7) Дополнительно на последнем слое уточняются области интересов, используя простой линейный регрессор. В результате чего области наиболее точно соответствуют интересующим объектам.

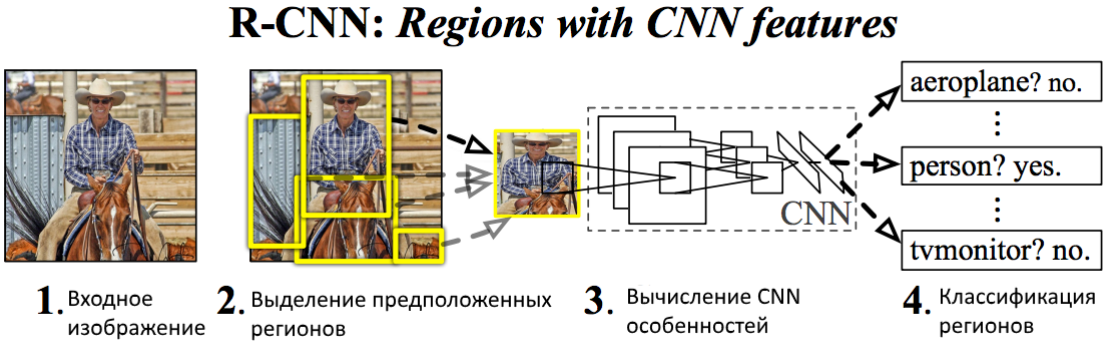


Рис. 7: Принцип работы R-CNN

Однако такой подход требовал очень большого объёма вычислений, решением проблемы выступила модифицированная версия этого подхода называемая Faster R-CNN.

**Faster R-CNN** Основными улучшениями, по сравнению с R-CNN стали:

- Объединение регионов интересов (RoI Pooling).
- Объединение этапов вычисления признаков, классификации и регрессии в одну модель.
- Отказ от процесса Selective search в обмен на использование CNN сети, которая и так находит признаки, только теперь использует их для предсказания регионов интересов.


Рис. 8: Принцип работы Faster-RCNN

Итогом стала модель(рис. 8) Особенностью архитектуры Faster-RCNN является двух-этапный подход. В котором на первом этапе выделяются регионы интересов, а на втором этапе классифицируются объекты, находящиеся в этих регионах. Также существует множество модификаций этой архитектуры, с использованием Alexnet, VGG, ResNet энкодеров, что добавляет гибкость при использовании этой архитектуры.

***AP архитектуры Faster R-CNN на датасете COCO составил 36.2***

### 2.2.2. SSD (Single Shot Multibox Detector)

Подход SSD[17] основан на сети прямого распространения (feed-forward network), которая создаёт фиксированное количество множеств ограничительных рамок (bounding box) и высчитывает уверенность принадлежности объектов в этой рамке к определённой категории. Следующим шагом осуществляется подавление немаксимумов (Non-Maximum Suppression), который помогает получить финальные детекции(области с объектами внутри).

На начальных слоях сети используется стандартная архитектура для высококачественной классификации изображений (за исключением самих слоёв классификации), это называется "базовая сеть". Затем добавляется вспомогательная структура для создания детекций с ключевыми особенностями.

**Мультиразмерная карта признаков для детекции** В конец базовой сети добавляются свёрточные слои признаков. Они позволяют произвести классификацию детекций разных масштабов.

**Предсказание детекиций** Каждый слой признаков делает фиксированное количество предсказаний на задетектированных областях, полученных с предыдущих слоёв.

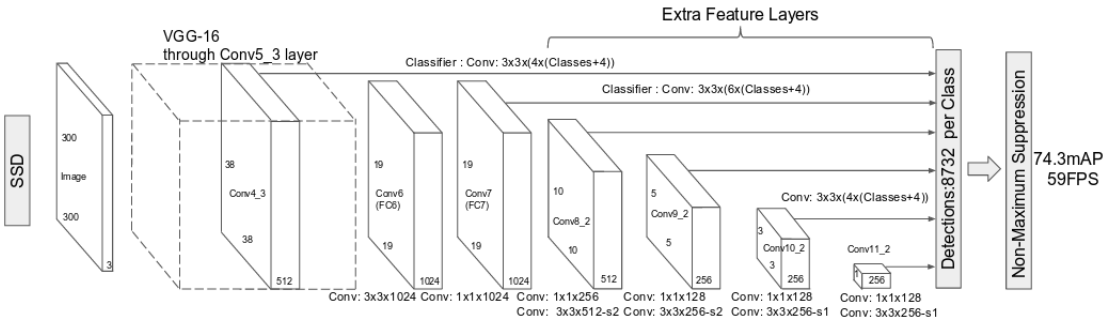


Рис. 9: Схема SSD

По сути архитектура SSD использует промежуточные слои нейронной сети, чтобы создать карту признаков, на которую потом накладываются фильтры для классификации и предсказаний границ объектов. SSD является ярким представителем одноэтапных детекторов, что позволяет достичь высокой производительности, однако сказывается на качестве детектирования и классификации объектов.

***AP архитектуры SSD на датасете COCO составил 31.2***

### 2.2.3. YOLO (You Only Look Once)

Подход данной модели заключается в применении единственной нейронной сети ко всему изображению. Эта сеть разбивает изображение на регионы, предсказывает ограничительные рамки и вероятности для каждой из них. Эти ограничительные рамки являются взвешенными, основываясь на вероятности предсказания.

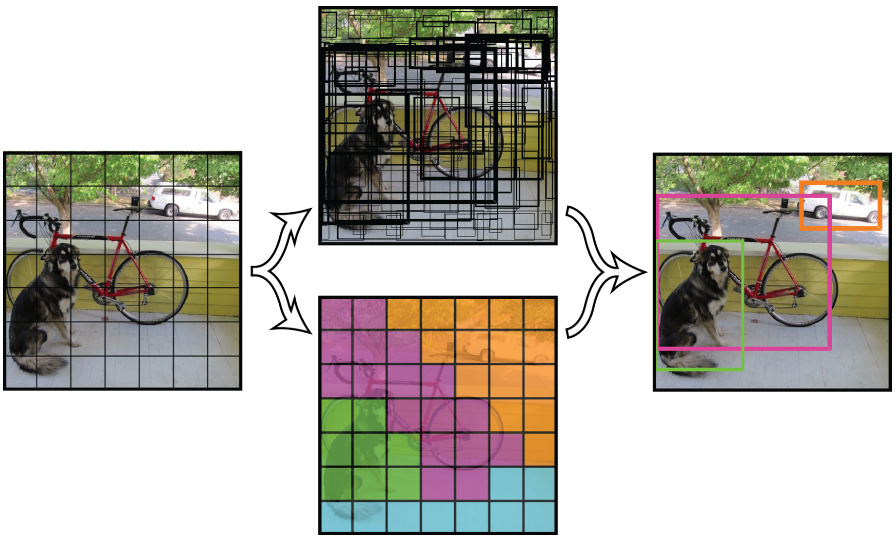

Рис. 10: принцип работы YOLO

Схема строения YOLO[22] больше напоминает SSD, нежели R-CNN. Ключевыми особенностями являются полностью связанные слои, используемые в конце, размер матрицы входного изображения, а также способ объединения признаков (pooling). Производительность YOLO меньше по сравнению с SSD, однако была выпущена версия YOLOv2, в которой автор использовал "несколько хитростей" для увеличения производительности и точности. YOLOv2 так же является одноэтапным

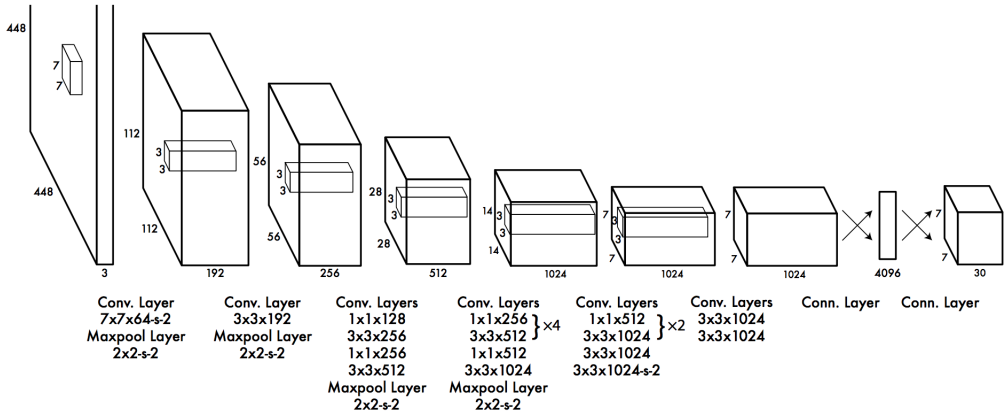


Рис. 11: Схема YOLO

детектором со всеми вытекающими преимуществами и недостатками. ***AP архитектуры YOLOv2 на датасете COCO составил 21.6***

### 2.2.4. Выбор

В данной задаче извлечения модели из изображения главную роль играет точность детектирования объекта на изображении, а не скорость работы самого решения. В связи с чем предпочтение отдаётся Faster-RCNN, который в силу того, что является двухэтапным детектором, позволяет более гибко с собой работать.

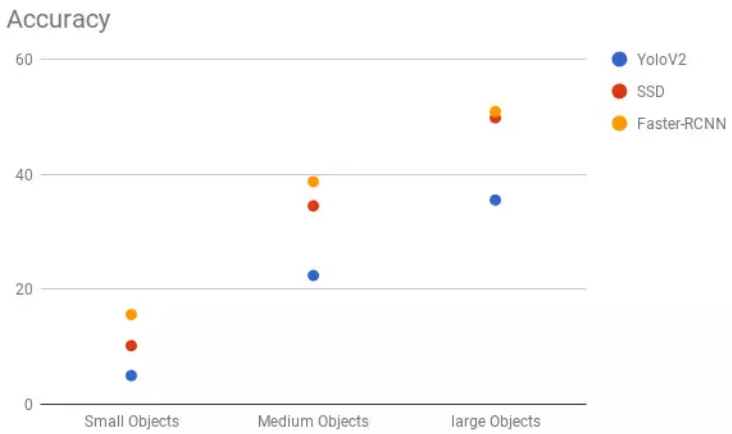


Рис. 12: YOLOv2 vs SSD vs R-CNN

## 2.3. Реконструкция трёхмерной модели

Для оценки качества работы алгоритмов по извлечению трёхмерной модели будет использоваться за метрику будет взято Chamfer distance, которое характеризует схожесть двух облаков точек $P_1, P_2 \subset \mathbb{R}$

$$CD(P_1, P_2) = \frac{1}{|P_1|} \sum_{x \in P_1} \min_{y \in P_2} \|x - y\|_2 + \frac{1}{|P_2|} \sum_{x \in P_2} \min_{y \in P_1} \|x - y\|_2 \qquad (2)$$

Для каждой точки в одном множестве ищется ближайшая точка из другого множества, а затем высчитывается среднее по все расстояниям.

В своём большинстве подходы представляют T-образные архитектуры типа (encoder-decoder) для обучения моделей и L-образные архитектуры для их тестирования[11]. Во время обучения происходит процесс минимизации двух функций потерь, одна функция отвечает за соответствие трёхмерной модели её скрытому представлению, другая функция отвечает за соответствие скрытого представления трёхмерной модели скрытому представлению, полученному из двумерного изображения.(рис. 13)

Для сравнения имеет смысл рассмотреть два алгоритма 3D-R2N2[1], как самый популярный и AtlasNet[2], как самый прогрессивный на момент начала исследования.

### 2.3.1. 3D-R2N2: 3D Recurrent Reconstruction Neural Network

Алгоритм, особенностью которого является то, что в данном подходе нейронная сеть принимает на вход одно или несколько изображений объекта сделанных с разных углов обзора и возвращает реконструированный объект в виде набора вокселей[19], которые представляют из себя трёхмерную модель.

Одной из ключевых особенностей 3D-R2N2 является выборочное обновление скрытых представлений, что позволяет во время процесса тренировки адаптивно изучать информацию о трёхмерном представлении объекта, если существует несколько различных точек обзора одного и того же объекта(рис. 14).

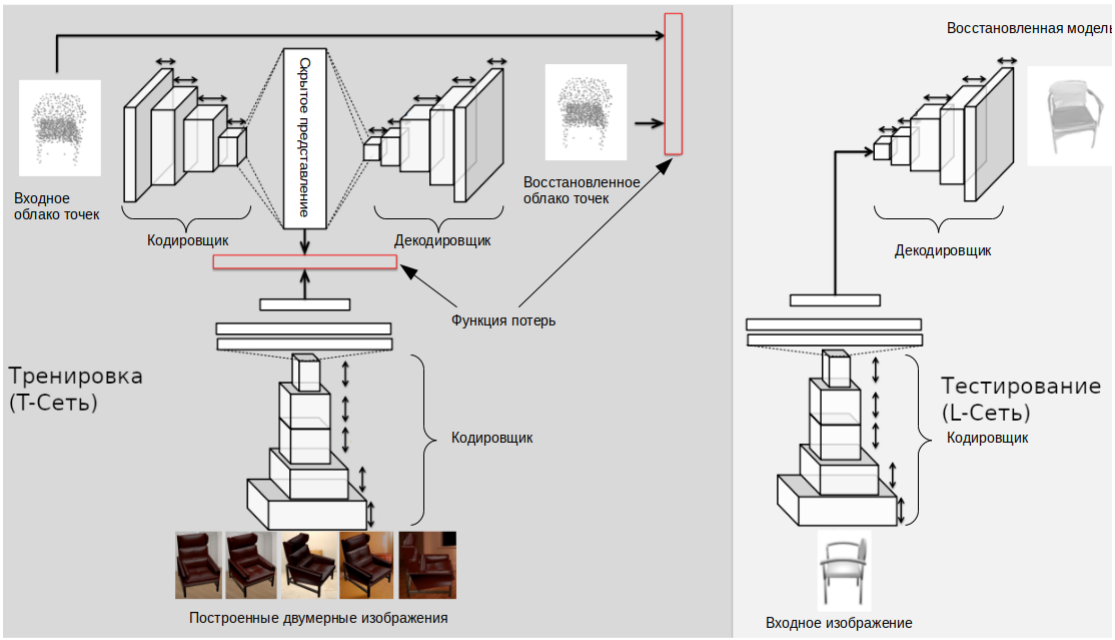


Рис. 13: Общий вид

3D-R2N2 представляет из себя encoder-decoder архитектуру с модулем долгой краткосрочной памяти посредине, где роль энкодера выполняет свёрточная нейронная сеть с остаточными связями (рис. 15). Выходной сигнал энкодера затем сглаживается и передается на полностью связный слой, который сжимает выходной сигнал в 1024-размерный вектор признаков.

Центральная часть архитектуры 3D-R2N2 является рекуррентным модулем, который позволяет сохранять в памяти увиденную информацию и обновлять её по мере поступления новых данных. Эта часть ответственна за одновременную работу однопозиционного и многопозиционного решения.

Декодер представляет из себя деконволюционную(обратную свёрточной) нейронную сеть, которая получает сигнал из рекуррентного модуля, а затем увеличивает его разрешение применением трёхмерных свёрток(3D convolutions), активационных функций, и 3D unpooling, до тех пор, пока сигнал не достигнет необходимого разрешения.

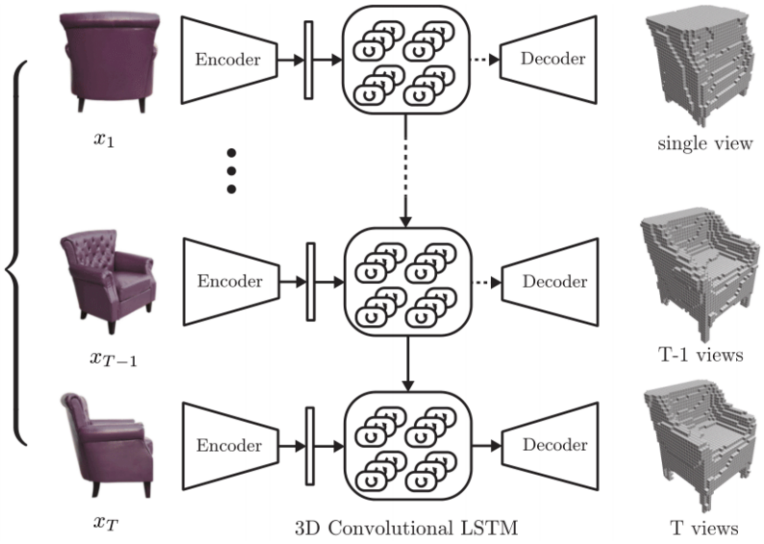


Рис. 14: Особенность 3D-R2N2

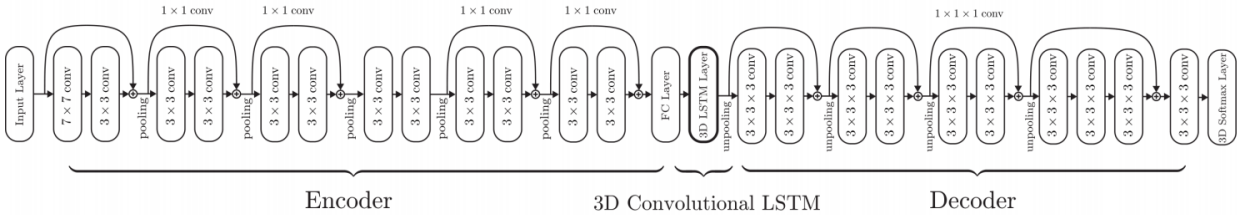


Рис. 15: Архитектура 3D-R2N2

В качестве функции потерь выступает softmax функция, применяемая к каждому вокселю $(i, j, k)$

$$L(\mathcal{X}, y) = \sum_{i,j,k} y_{(i,j,k)} \log(p_{(i,j,k)}) + (1 - y_{(i,j,k)}) \log(1 - p_{(i,j,k)}) \tag{3}$$

***Значение метрики Chamfer Distance при использовании данного подхода на датасете Pix3D составило 0.239***

### 2.3.2. AtlasNet: A Papier-Mache Approach to Learning 3D Surface Generation

Алгоритм, главных особенностей данного подхода является представление объекта не в виде набора вокселей, а в виде полигональной сетки(меша), что упрощает восприятие формы глазом, а также является более продвинутым подходом. Авторы алгоритма представляют

поверхность как топологическое пространство, которое локально напоминает Евклидову поверхность. Попыткой подхода является локальная аппроксимация целевой поверхности, путём отображения на неё множества прямоугольных элементов. Использование множества таких элементов позволяет моделировать сложные поверхности. Подобный подход применяется в технологии папье-маше, откуда и название у алгоритма.

Для алгоритма поставлено две задачи

- Кодировка и декодировка трёхмерного объекта, из полученных данных в виде облака точек
- Восстановление трёхмерной формы объекта, из полученных данных в виде фотографии

За основу алгоритма взята архитектура PointNet[16], которая выступает в качестве энкодера для набора точек. Этот энкодер преобразовывает облако точек в вектор размерности $k = 1024$ (рис. 16) Затем сигнал подаётся в декодер, который представляет из себя 4 полносвязных слоя с размерностями 1024, 512, 256, 128 с линейными выпрямителями (ReLU) на первых трёх слоях и гиперболической тангенциальной функцией активации на последнем слое (tanh). На выходе алгоритма возвращается облако точек восстановленного объекта, между которыми "натягиваются" полигоны.

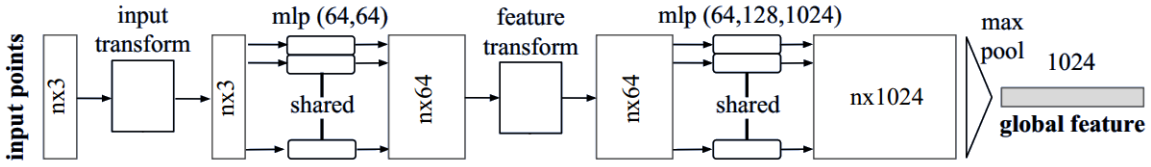


Рис. 16: Архитектура энкодера PointNet

В качестве функции потерь используется Chamfer Distance, значение которой минимизируется для достижения большего соответствия между оригинальной моделью(Ground Truth) и её реконструированной версией.

***Значение метрики Chamfer Distance при использовании данного подхода на датасете Pix3D составило 0.126***

### 2.3.3. Выбор

В связи с меньшим значением метрики (меньше-лучше), а так же более прогрессивным подходом к реконструкции модели объекта, было решено выбрать архитектуру AtlasNet, в качестве решения для SVR. Так же AtlasNet использует архитектуру ResNet в качестве энкодера для изображения, что будет полезным в дальнейшем.

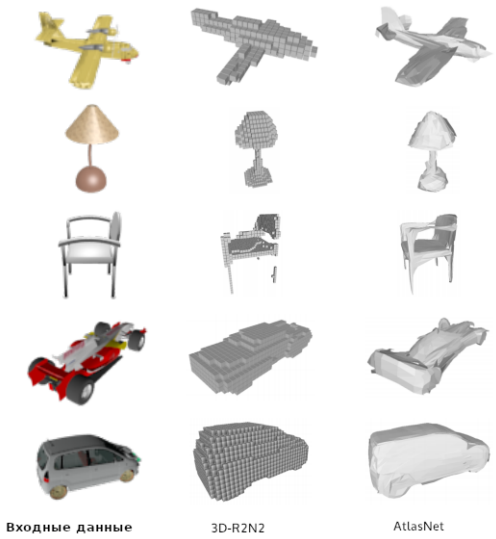


Рис. 17: Сравнение результатов работы AtlasNet и 3D-R2N2

# 3. Глава 3. Разработка

## 3.1. Проектирование

Основной идеей была реализация подхода, позволяющего детектировать объекты на изображении реального мира, а затем производить их реконструкцию (рис. 18).

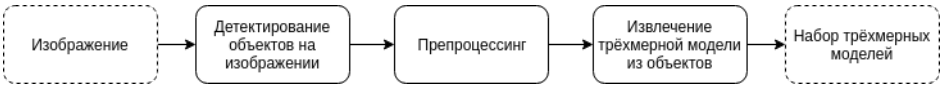


Рис. 18: План

В качестве детектора был выбран Faster-RCNN, а в качестве решения для SVR был выбран AtlasNet. Для реализации поставленного плана достаточно взять L-образную часть сети с весами, натренированными на необходимых данных. На вход же этой сети подать набор изображений, являющимися вырезанными участками исходного изображения, предоставленные детектором.

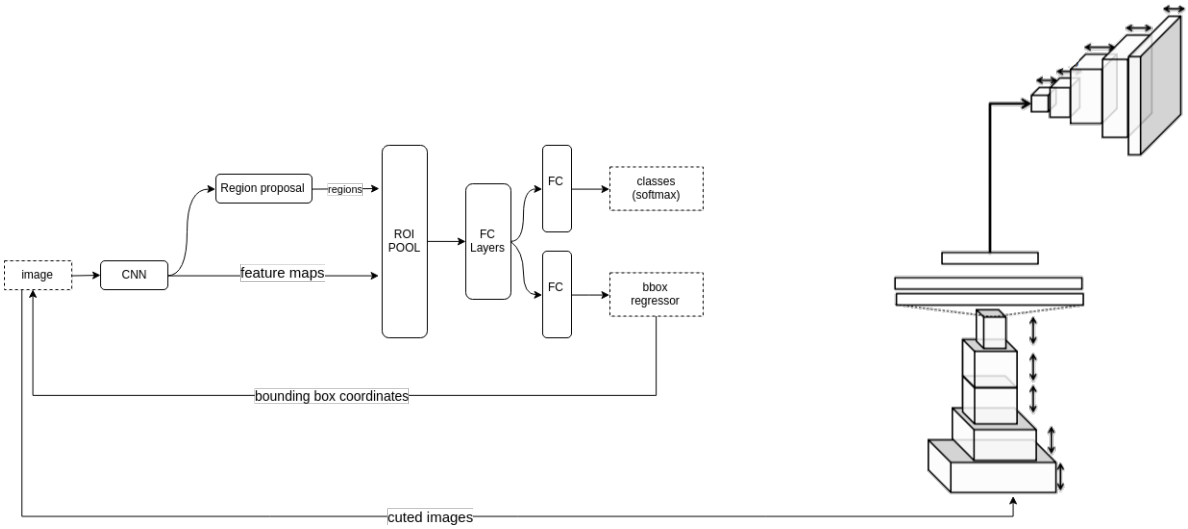


Рис. 19: Подход ”в лоб”. В левой части представлена архитектура Faster-RCNN. В правой части L-образная часть SVR сети.

Такой подход является полностью рабочим, однако крайне затратным в плане вычислений и не оптимальным в архитектурном плане. Так как участки интересов(regions of interests, ROI) кодируются два раза. В первый раз энкодером Faster-RCNN(в составе целого изображения), а затем энкодером Atlasnet(как отдельные куски изображения). Принимая во внимание тот факт, что в обоих случаях используется архитектура ResNet, предобученная на ImageNet[10], то и особенности(features) извлекаются одни и те же на каждом этапе. Решением проблемы является отсечение энкодера в L-образной части и передача регионов интересов вместе с выделенными особенностями напрямую в декодер L-образной части.

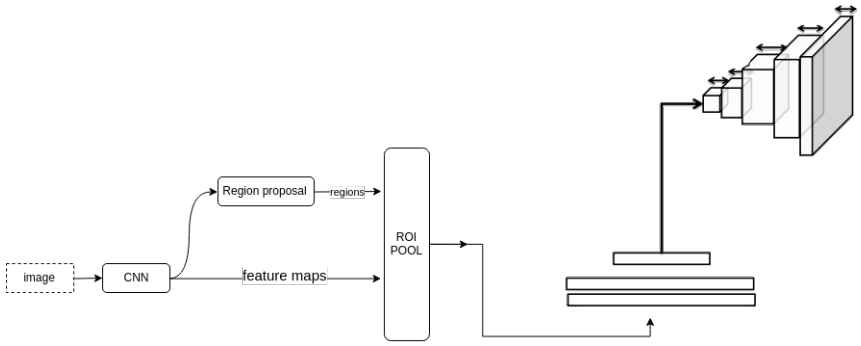


Рис. 20: Подход с передачей сигнала с регионами интересов напрямую в декодер L-образной части

## 3.2. Реализация

В качестве языка программирования был выбран python, так как это самый популярный язык, на котором проводятся исследования в области глубокого обучения. В качестве библиотеки для работы с методами глубокого обучения был выбран PyTorch, в силу своей гибкости и простоты отладки исполняемого решения. Так как модификации задействуют только тестовую часть архитектуры, можно воспользоваться предобученными весами.

Для начала была взята реализация алгоритма Faster-RCNN[7]. Из которой был получен набор векторов признаков для каждой интересу-

ющей области. Этот массив данных представляет собой, массив элементов размерности $(n \times 1024)$, где n - количество областей интересов.

Следующим шагом являлась модификация тестовой части AtlasNet. Была взята реализация[3], из которой извлеклась часть, отвечающая за декодер. Функция `forward_inference_from_latent_space` осуществляет восстановление трёхмерной модели, получая на вход вектор признаков.

Listing 1: Восстановление 3D модели

```
def forward_inference_from_latent_space(self, x, grid):
    outs = []
    for i in range(0, self.nb_primitives):
        rand_grid = Variable(torch.cuda.FloatTensor(grid[i]))
        rand_grid = rand_grid.transpose(0, 1).contiguous().unsqueeze(0)
        rand_grid = rand_grid.expand(x.size(0), rand_grid.size(1),
            rand_grid.size(2)).contiguous()
        y = x.unsqueeze(2).expand(x.size(0), x.size(1),
            rand_grid.size(2)).contiguous()
        y = torch.cat( (rand_grid, y), 1).contiguous()
        outs.append(self.decoder[i](y))
    return torch.cat(outs, 2).contiguous().transpose(2,1).contiguous()
```

В эту функцию итеративно подавались элементы массива, полученного из модифицированной части Faster-RCNN. На выходе получалось n восстановленных трёхмерных моделей для каждого региона интересов.

## 3.3. Результаты

Разработанное решение проверялось на реальных данных, однако так как получить размеченные данные объектов реального мира - далеко не тривиальная задача, то оценка результатов работы алгоритма будет проводиться визуально. Оценка же работы отельных частей алгоритма представлена в пунктах 4.3.2(на синтетических данных), 4.2.1(на изображениях реального мира)

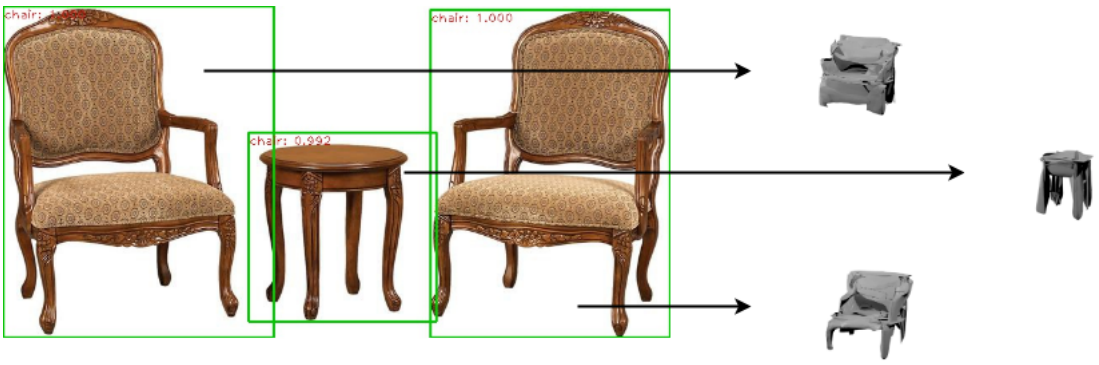

Рис. 21: Решение отработало хорошо. На изображении были определены объекты для реконструкции и произведена их трёхмерная реконструкция.

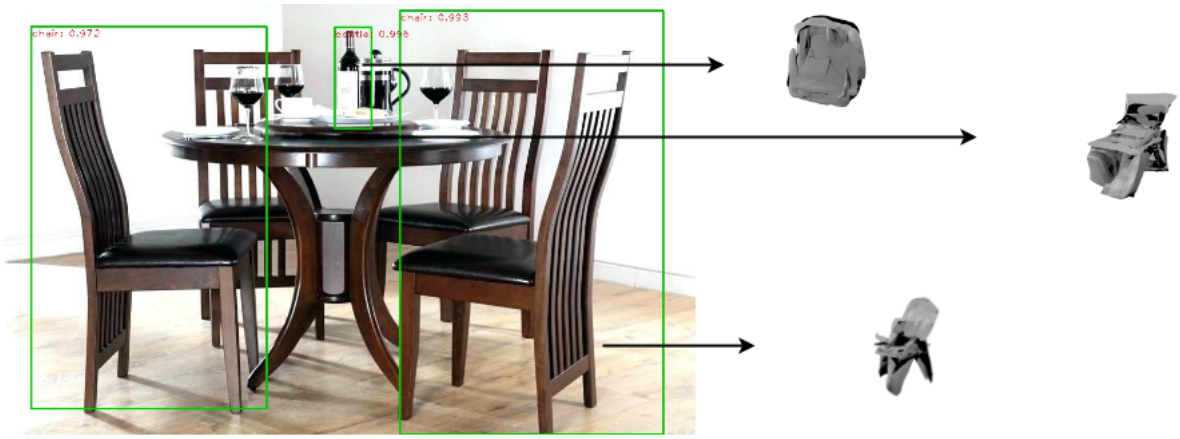

Рис. 22: Решение отработало не лучшим образом. На изображении были определены объекты для реконструкции, однако алгоритм не смог реконструировать бутылку на столе. Кроме того реконструкция стульев была произведена с большими ошибками.

### 3.3.1. Метрика

В качестве измерения результатов работы алгоритма была введена следующая метрика. Semblance of Extraction

$$SE(O_1, ..., O_n) = \frac{1}{n} \sum_{n} SF(O_i) \tag{4}$$

- $O_i$ - Отвечает за извлечённый объект. Так как оценка производилась визуально, то и представление объекта бралось визуальное.
- $n$ - Количество извлечённых объектов.

- $SF$ - Similarity Function, функция, значение которой изменяется в пределах от 0 до 1. и определяется пользователем, где 1 - полное сходство объекта с представленным на изображении и 0 - полное несоответствие.

Метрика характеризует среднее качество извлечения трёхмерной модели, определённое пользователем "на глаз". К такому подходу пришлось прибегнуть, потому что нет открытых датасетов с фотореалистично сгенерированными моделями и построенными сложными сценами из них(частичная видимость пересечения и так далее).
***Значение метрики на тестовой выборке составило*** $\approx 0.3$
(Так как достичь результатов в районе 1-0.8 - непосильная задача для современных алгоритмов, полученную точность можно считать относительно приемлемой.)

# Выводы

- Был проведён теоретический анализ предметной области. В результате которого было принято решение использовать методы глубокого обучения для достижения поставленной цели.
- Был проведён анализ и сравнение существующих решений как в области детектирования объектов на изображении, так и в области извлечения трёхмерных моделей из этих объектов.
- Была разработана архитектура для решения поставленной задачи, а так же реализован программный продукт, этой архитектуре соответствующий.
- Было проведено тестирование разработанного программного продукта.

# Заключение

Разработанное решение выполняет поставленные задачи, однако оно очень чувствительно к фону извлекаемого объекта, а так же к пересечениям извлекаемого объекта с другими объектами. Это накладывает ограничения на область его применения. Проблема кроется в архитектурных особенностях сети AtlasNet, а так же в типе данных, на которых эта сеть обучалась. Изображения, подаваемые на вход при обучении были "идеальной версией" трёхмерных объектов, которые они представляли, с простым фоном, без наложений других объектов, под определёнными углами и так далее. Решением этой проблемы является модификация принципа генерации двумерных представлений трёхмерных объектов в угоду фотореалистичному качеству и построениям сложных сцен из реального мира, а так же внедрение дополнительного модуля, выполняющего семантическую сегментацию в процесс тестирования, для того чтобы с более высокой точностью определять границы объекта и тем самым увеличить точность реконструкции.

# Список литературы


[1] 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction / Christopher B Choy, Danfei Xu, JunYoung Gwak et al. // Proceedings of the European Conference on Computer Vision (ECCV). — 2016.

[2] AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation / Thibault Groueix, Matthew Fisher, Vladimir G. Kim et al. // CoRR. — 2018. — Vol. abs/1802.05384. — 1802.05384.

[3] AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation / Thibault Groueix, Matthew Fisher, Vladimir G. Kim et al. // Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). — 2018.

[4] Belhumeur Peter, Kriegman David, Yuille A.L. The bas-relief ambiguity. — Vol. 3. — 1997. — 01. — P. 1060–1066.

[5] Brooks. M. Two results concerning ambiguity in shape from shading. // AAAI-83. — 1983. — P. 36–39.

[6] Estellers V., Schmidt F., Cremers D. Robust Fitting of Subdivision Surfaces for Smooth Shape Analysis // Proc. of the Int. Conference on 3D Vision (3DV). — 2018. — September.

[7] A Faster Pytorch Implementation of Faster R-CNN / Jianwei Yang, Jiasen Lu, Dhruv Batra, Devi Parikh // https://github.com/jwyang/faster-rcnn.pytorch. — 2017.

[8] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun // CoRR. — 2015. — Vol. abs/1506.01497. — 1506.01497.

[9] Hartley R. I., Zisserman A. Multiple View Geometry in Computer Vision. — Second edition. — Cambridge University Press, ISBN: 0521540518, 2004.

[10] ImageNet: A Large-Scale Hierarchical Image Database / J. Deng, W. Dong, R. Socher et al. // CVPR09. — 2009.

[11] Learning a Predictable and Generative Vector Representation for Objects / Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, Abhinav Gupta // CoRR. — 2016. — Vol. abs/1603.08637. — 1603.08637.

[12] Motion Cooperation: Smooth Piece-Wise Rigid Scene Flow from RGB-D Images / M. Jaimez, M. Souiai, J. Stueckler et al. // Proc. of the Int. Conference on 3D Vision (3DV). — 2015. — .

[13] ObjectNet3D: A Large Scale Database for 3D Object Recognition / Yu Xiang, Wonhui Kim, Wei Chen et al. // European Conference Computer Vision (ECCV). — 2016.

[14] Oliensis. J. Shape from shading as a partially well–constrained problem. // CVGIP: Image Understanding. — 1991. — P. 54(2):163–183.

[15] Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling / Xingyuan Sun, Jiajun Wu, Xiuming Zhang et al. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2018.

[16] PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation / Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas // CoRR. — 2016. — Vol. abs/1612.00593. — 1612.00593.

[17] SSD: Single Shot MultiBox Detector / Wei Liu, Dragomir Anguelov, Dumitru Erhan et al. // CoRR. — 2015. — Vol. abs/1512.02325. — 1512.02325.

[18] ShapeNet: An Information-Rich 3D Model Repository / Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas et al. // CoRR. — 2015. — Vol. abs/1512.03012. — 1512.03012.

[19] Shin Daeyun, Fowlkes Charless C., Hoiem Derek. Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction // CoRR. — 2018. — Vol. abs/1804.06032. — 1804.06032.

[20] Synthesizing 3D Shapes via Modeling Multi-View Depth Maps and Silhouettes with Deep Generative Networks / Amir Arsalan Soltani, Haibin Huang, Jiajun Wu et al. // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2017. — P. 1511–1519.

[21] Xiang Yu, Mottaghi Roozbeh, Savarese Silvio. Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild // IEEE Winter Conference on Applications of Computer Vision (WACV). — 2014.

[22] You Only Look Once: Unified, Real-Time Object Detection / Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, Ali Farhadi // CoRR. — 2015. — Vol. abs/1506.02640. — 1506.02640.

[23] ichi Kanatani Ken, Chou Tsai-Chia. Shape from texture: General principle // Artificial Intelligence. — 1989. — Vol. 38, no. 1. — P. 1 – 48. — URL: http://www.sciencedirect.com/science/article/pii/0004370289900660.